Alice Liu, Emily Kim, Lily Liang

Professor Zhao Yunhua

CSC448 Artificial Intelligence

12.20.2023

Project Documentation

Problem

Addressing the challenge of differentiating between spam and non-spam emails is the main focus of our project's classification task. Even while current algorithms are able to categorize emails, spam continues to find its way into our inboxes, suggesting that further improvement is necessary. We focused on developing a classification system that is more accurate and reliable, reducing the amount of spam emails that reach an individual's user's inbox.

Data

To prepare our data for a model, a crucial step involves data cleaning. Initially, we identified 2 datasets: spam.csv and spam_or_not_spam.csv. Here are the specifics of each datasets:

- spam.csv
    - 2 columns, 5k rows
    - Columns: Category, Message
- spam_or_not_spam.csv
    - 2 columns, 3k rows
    - Columns: Email, Label

Both datasets exhibited similar formatting, prompting us to merge them and create a more extensive dataset for our model. During this combination, meticulous attention was paid to ensure uniformity in format. This involved relabeling columns, standardizing values to represent spam

indicators as 0 and 1, removing NULL values, and eliminating any duplications. Further insights into this data combining process are documented in our "Combining Data.ipynb". The resulting dataset is characterized by the following details:

- combined_data_clean.csv
    - 2 columns, 8k rows
    - Columns: Label, Email

Utilizing the combined CSV file, we conducted additional cleaning and data preprocessing to refine the dataset further. The ultimate dataset employed in our modeling is denoted below:

- preprocessed_english.csv
    - 3 columns, 8k rows
    - Columns: Label, Email, processed_email

Accomplished

The goal of our project was to improve the accuracy of our spam classification model. We were able to confirm the Random Forest model's fundamental function in our prototype by reaching an accuracy score of 97% and a precision of 96% using its model.

- **Model Exploration:** We explored six models – Naive Bayes (multinomial naive bayes), SVM, Logistic Regression, Random Forest, Gradient Boosting, and Decision Tree – each model selected based on how well it could handle particular email classification tasks.
- **Model Performance:** The Random Forest model demonstrated to be the most accurate and best precision compared to the other five models, achieving a 97% accuracy score and a 96% precision score.

- **User Interface Development:** We used the Voilà jupyter server extension to produce a working interface where users can test to see whether their text would be considered as spam or not.

Project Organization

These were the key notebooks to focus on:

1. Combining data: Combining Data.ipynb
2. Cleaning and Preprocessing: PREPROCESSING.ipynb
3. Final Dataset: preprocessed_english.csv
4. Model Building: Model Building.ipynb
5. Model Evaluation: Model Evaluation.ipynb
6. Viola: Voila Implementation.ipynb

Task Division

Task	Assigned
Create Github Repository	Everyone
Brainstorm Project	Everyone
Finding Dataset	Everyone
Combine data and removing duplication	Everyone
EDA	Everyone
Data Preprocessing	Everyone

Modeling: Gradient Boosting, Random Forest	Alice
Modeling: SVM, Logistic Regression	Emily
Modeling: Naive Bayes, Decision Tree	Lily
Model Evaluation	Alice, Lily
Create User Interface	Lily, Emily

NOTE: Everyone means each person in the team tried it on their own file, and we come together through zoom to discuss, combine, and make decisions. Moreover, everyone contributed and helped each other in each part of the task.